J Mol Evol (1984) 20:215–226

Journal of
Molecular Evolution


# Evolving Sea Urchin Histone Genes— Nucleotide Polymorphisms in the H4 Gene and Spacers of *Strongylocentrotus purpuratus*

Lawrence N. Yager, John F. Kaumeyer, and Eric S. Weinberg

Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

**Summary.** We present a comparison of spacer and coding sequences of histone gene repeats from four *Strongylocentrotus purpuratus* individuals. Sequences of two previously cloned units (pCO2 and pSp2) were compared with three new histone gene clones, two of them from a single individual. Within a 1.7-kb region, 59 polymorphic sites were found in spacers, in mRNA nontranslated stretches, and at silent sites in codons of the H4 gene. The permitted silent-site changes were as frequent as in any other region studied. The most abundant polymorphisms were single-base substitutions. The ratio of transitions : tranversions : single-base-pair insertions/deletions was 3:2:2. A number of larger insertions/deletions were found, as well as differences in the length of $(CTA)_n$ and $(CT)_n$ runs. Two of the five cloned repeats contained an insertion of a 195-bp element that is also present at many other sites in the genomes of every *S. purpuratus* individual studied. Pairwise comparisons of the different clones indicate that the variation is not uniformly divergent, but ranges from a difference of 0.34% to 3.0% of all nucleotide sites. A parsimonious tree of ancestry constructed from the pairwise comparisons indicates that recombination between the most distantly related repeats has not occurred in the 1–2 million years necessary for accumulation of the variation. The level of sequence variation found within the *S. purpuratus* population, for both tandemly repeated and single-copy genes, is 25%–50% of that found between *S. purpuratus* and *S. drobachiensis*.

**Key words:** Histone genes — Polymorphisms — DNA sequence — Tandemly linked genes — Haplotypic tree — Silent substitutions — Transitions/transversions — Insertions/deletions — Spontaneous mutation

*Offprint requests to:* E.S. Weinberg

## Introduction

DNA sequence analysis provides a powerful method for investigation of the phylogenetic relationships of organisms, the divergence of multigene families, and the evolution of gene structure. Few studies have compared actual DNA sequences of alleles of a particular gene from a natural population, even though such studies are essential if we are to learn the nature and extent of true genetic variation. Although comparison of restriction enzyme recognition sites can give a rough estimate of gene divergence, only direct DNA sequence analysis provides definitive information on the sites of polymorphism and the nature of the mutational events.

The histone genes of the sea urchin constitute a multigene family (reviewed most recently by Maxson et al. 1983a). Although there are several histone gene sets in the sea urchin genome, the genes discussed here code for the "early" embryonic H4 histone. The genes coding for "late" embryonic histones have a very different sequence and a completely different gene organization (Childs et al. 1982; Maxson et al. 1983b; J.F. Kaumeyer and E.S. Weinberg, manuscript in preparation) and will not be discussed. Figure 1A presents a map of the early embryonic histone gene repeat. A unit of 6–7 kb contains sequences coding for the five histone mRNAs; these genes are separated from one another by nontranscribed spacer DNA. This unit is represented several hundred times per haploid genome, mostly in a cluster of tandem repeats.

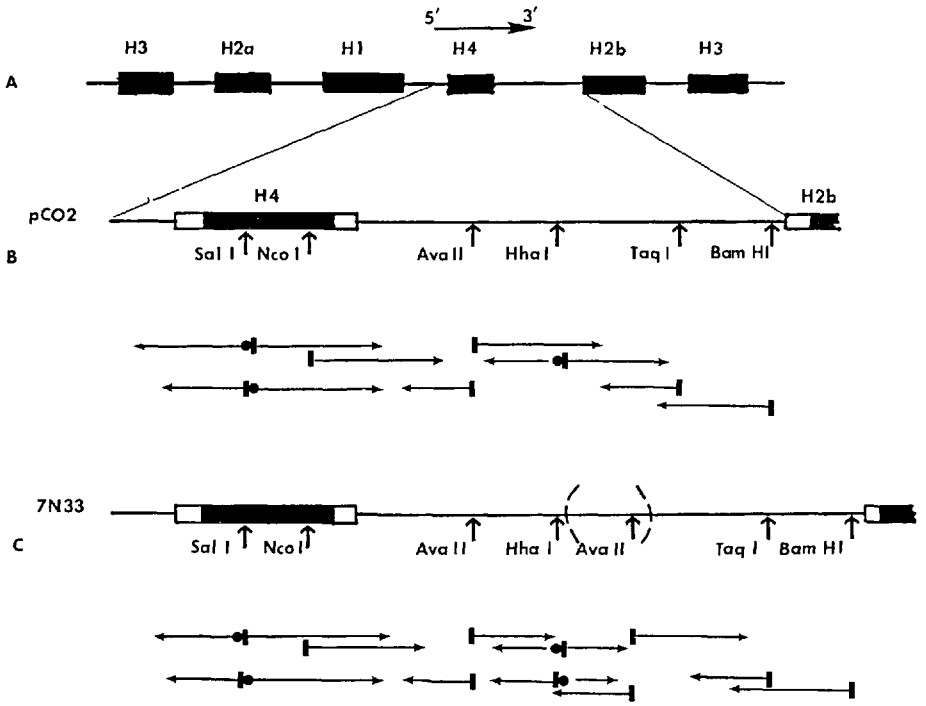


**Fig. 1A-C.** Map of the histone repeat unit and sequencing strategy. **A** Organization of the repeat unit. **B** Enlargement of the H4 gene region and flanking spacers of pCO2. The restriction sites found in both pCO2 and 6S35 are indicated; the sequencing strategies were identical for both clones. Individual labeling reactions and sequence runs for pCO2 are shown below the map. The two upper lines of arrows represent one DNA strand; the two lower lines, the other strand. Fragments labeled at the 3′ end with the Klenow fragment of DNA polymerase I, reverse transcriptase, or T4 polymerase are marked with vertical bars; those labeled at the 5′ end with polynucleotide kinase are indicated by filled circles. Arrows indicate the direction of sequencing from the labeled end. **C** The H4 gene region and flanking spacers of 7N33. The sequencing strategies for clones 6S46 and 7N33 were identical; the labeling reactions for 7N33 are shown. The 195-bp insert is indicated by the dashed parentheses

Two major types of repeat unit usually occur within any individual sea urchin; ostensibly each type of repeat is inherited from a different parent. Some individuals, however, contain three major repeat units of different sizes. The types differ from one another in repeat length and in restriction enzyme maps (Overton and Weinberg 1978), but the repeat units of a single type are highly homogeneous. Until now no actual DNA sequence information has been available to gauge the similarity or divergence of repeats either from a single cluster of repeats or from clusters from different individuals. Since all evidence indicates that very large clusters containing one or two main repeat types are the predominant sites of the genes, we regard the clusters as alleles and the sequences presented below as the typical repeats of that cluster.

In the work presented here, we compared sequences of five copies of the sea urchin early embryonic H4 histone gene and the surrounding spacer DNA. Two copies were derived from a single individual, and the other three copies were cloned from three other individuals. We report here the extent and locations of sequence variation as well as the nature of the mutational events. The data allow the construction of a lineage of divergence of the five copies and indicate that there has been no recombination between the genes of the two observed basic sequence patterns.

The results provide the first comparison of repeat sequence data from different individuals and different clusters to illustrate the extent of histone gene sequence divergence within a *Strongylocentrotus purpuratus* population.

## Materials and Methods

*Preparation and Restriction Enzyme Digestion of Sea Urchin DNA.* DNA from the sperm of single *S. purpuratus* individuals, which we denote as 6S (obtained from Pacific Biomarine, Inc., Venice, California) and 7N (obtained from Peninsula Marine Biologicals, Sand City, California), was prepared by methods slightly modified from those of Stafford and Bieber (1975) and Joseph and Stafford (1976). Restriction enzymes (New England Biolabs or Bethesda Research Laboratories) were used as recommended by the suppliers.

*Preparation and Screening of Sea Urchin Libraries.* DNA libraries were prepared from individual sea urchins by the method of Blattner et al. (1977), using the in vitro packaging protocol of Hohn and Murray (1977). Genomic DNA was digested with Sac I (which recognizes only one site within the *S. purpuratus* histone gene repeat unit) and cloned into the Sac I sites of the vector λgt.λB (Thomas et al. 1974). Clones containing histone DNA

were identified by hybridization to filter blots (Benton and Davis 1977) with nick-translated pCO2A DNA. [pCO2A is a complete histone gene repeat unit, first cloned into the Hind III site of pBR313, and now recloned into pBR322 (Overton and Weinberg 1978).] Three phage clones were selected for further study: two from individual 6S and one from the 7N library. A Bam HI fragment containing the complete H4 and H1 genes, part of the H2A gene, and some phage DNA was subcloned from each phage into pBR322. The three subclones, designated 6S35, 6S46, and 7N33, were used for sequencing. Additional details of the cloning will be provided elsewhere. The other clone that we sequenced was pCO2A, obtained from another sea urchin individual some years ago (Overton and Weinberg 1978).

*DNA Sequencing, Sequencing Strategy, and Computer Analysis.* DNA fragments were labeled either at the 5′ end using polynucleotide kinase (Maxam and Gilbert 1980) or at the 3′ end using the Klenow fragment of DNA polymerase I or reverse transcriptase. For those enzymes that generate 3′ overhangs, the 3′ end was labeled using the T4 DNA polymerase exchange reaction described by Maniatis et al. (1982). DNA fragments were isolated from polyacrylamide gels by electroelution and purified on small columns of Whatman DE53 cellulose. DNA sequencing was done by methods slightly modified from those of Maxam and Gilbert (1980).

The strategy used to sequence the clones is shown in Fig. 1. Two of the four clones contain a 195-bp element that includes an Ava II site. The fragments labeled for sequencing were therefore somewhat different for clones with and without the insert; only one example of each type is presented in Fig. 1B and C. DNA sequences were analyzed using the SEQ programs described by Brutlag et al. (1982).

## Results

### *Comparison of Sequences*

The H4–H2B spacer sequence was chosen for analysis because we had previously found length and sequence polymorphisms in this region (Overton and Weinberg 1978). The most striking difference is an insert of 195 bp that we found in 6 of 22 individual sea urchins. This insert is a member of a family of intermediate repetitive sequences, and is present at many other sites in the genome (L.N. Yager, J.F. Kaumeyer, and E.S. Weinberg, manuscript in preparation). To compare the natures, extents, and sites of other mutational events, we sequenced the H4 gene and the surrounding region extending almost completely across the H4–H2B spacer to just 5′ to the H2B mRNA TATA box (Fig. 1). The sequence of approximately two-thirds of this region, derived from the pSp2 clone from the laboratory of L. Kedes, had already been published (Sures et al. 1978, 1980; Grunstein et al. 1981; Mauron et al. 1981). We sequenced four additional repeats: the previously cloned and characterized pCO2 (Overton and Weinberg 1978), two repeats representing two major repeat classes of a single individual sea urchin (6S46 and 6S35), and a major repeat from another individual (7N33). In total, five sequences from four different individuals were compared. Two of these sequences, 6S46 and 7N33, contained the 195-bp insert referred to above.

The sequence of pCO2 in the region of interest is shown in Fig. 2, starting 196 bp upstream from the ATG initiation codon of the H4 gene. A TATA-type sequence, TAACAATA in this case, is at −103, and the cap site is at −70 (Sures et al. 1980). The H4 coding sequence is 306 bp long and is followed by a 3′ nontranslated sequence that probably terminates at base 370 within a region of dyad symmetry (Hentschel and Birnstiel 1981). The H4–H2B spacer of pCO2 extends to the end of the region that we sequenced. The H2B TATA sequence is 10 bp further downstream from this point (i.e., at base 1385). The sequences for the three other repeats in the same region started at −193, −144, and −111 for 6S46, 6S35, and 7N33, respectively. The sequence between bases 430 and 850 has not previously been published for pSp2 or any other repeat. The H4–H2B spacer is an AT-rich sequence, with the newly obtained sequence particularly AT-rich (68%). In addition to the $(CTA)_7$ sequence at bases 1036–1056, there are regions of runs of other simple sequences, for example, $(CT)_5$ at 321–330, $(AAAG)_2T(AAAG)_2$ at 413–429, and $(AT)_3T(AT)_2$ at 858–868. In other areas, less perfect copies of a motif are found, as is often the case in nontranscribed DNA sequences. When the Intelligenetics SEQ program (Brutlag et al. 1982) was used to search for regions of longer direct or inverted repeat sequences, only one significant case was found: a 47-bp sequence at 491–537 homologous to a 45-bp sequence at 564–608, with only three single-base-pair mismatches and four single-base-pair gaps. This spacer sequence appears to have been created by sequence duplication.

### *Extent of Sequence Variation*

When the five sequences are compared, 59 sites of variation are found. The pCO2 sequence in Fig. 2 is keyed to show these 59 changes: A particular mark is placed at that point in the pCO2 sequence at which a difference is found in one or more of the other three examples that we sequenced or in pSp2. Of the 59 sites containing a difference in sequence in at least one repeat, 48 involve simple base changes and 2 involve changes in two consecutive bases. At one of the sites (no. 19) there is a base substitution in one repeat and a single-base insertion/deletion in another repeat; one of the larger insertion/deletion sites (site A) also represents two different events. The 59 sites noted therefore represent at least 61 different events. Figure 3 illustrates the actual changes. At each position that varies between any of the five examples, the bases of all five sequences are given. The positions and specific sequences of the larger insertions/deletions are given in Fig. 4A.

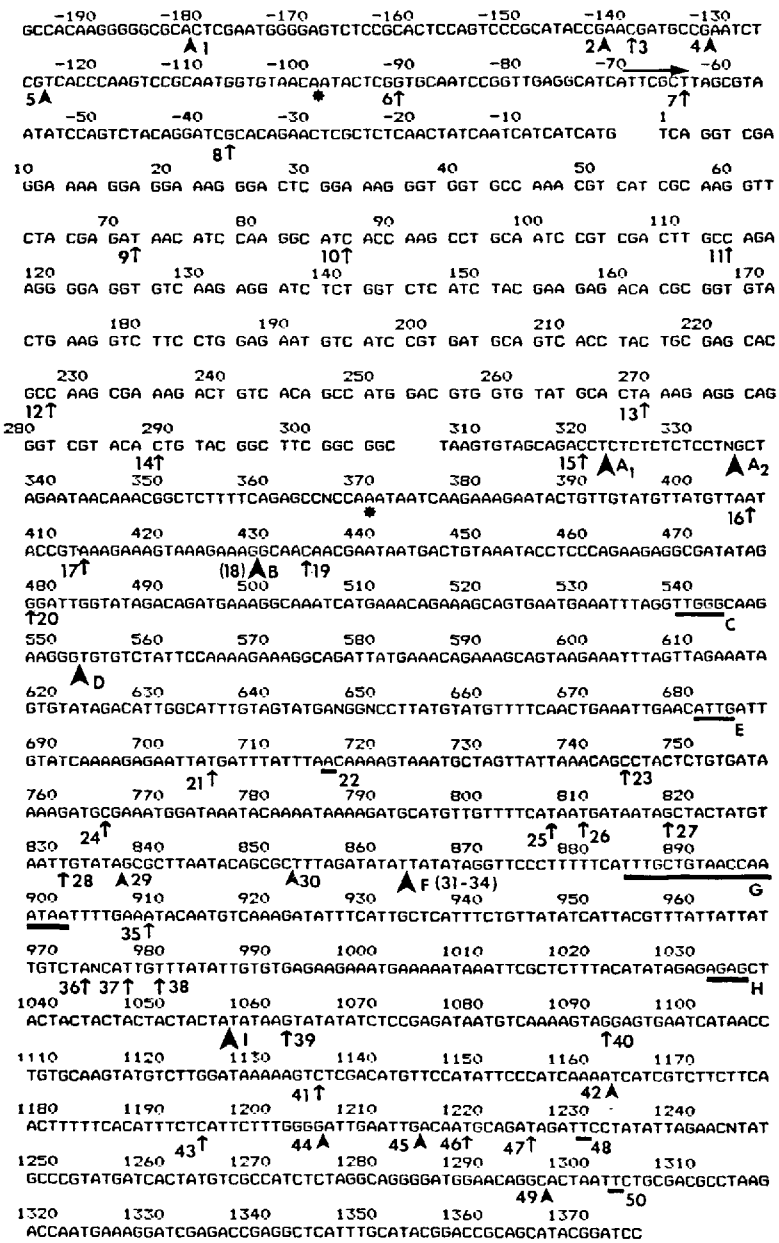

Fig. 2. Sequence of pCO2. Nucleotide positions that differ in any of the other repeats are indicated by numbers (1–50) or letters (A–I). Arrows represent positions at which base substitutions are found. Positions at which single-base insertions are found in another repeat are indicated by small numbered arrowheads. A position at which a base in pCO2 is not found in another repeat is indicated by an underline. Insertions/deletions larger than two bases are indicated by large lettered arrowheads and longer underlined stretches, respectively

The greatest difference in sequence is found between the pSp2 and 6S35 sequences. If each insertion/deletion is regarded as equal to a base change, these sequences differ in 31, or 3.0%, of the 1029 bases compared. Of the other pairwise comparisons made between sequences (over a longer region), the maximum difference is 2.7% for 6S35 versus pCO2. The least difference is found between 7N33 and 6S46, the two repeats that contain the inserted intermediate repetitive element. These sequences differ in only 6 positions (4 of which are within the 195-bp insert), or 0.34%, of the 1711 bp compared. In total, one can make ten pairwise comparisons between these five sequences. Figure 5 indicates the extent of variation between any two sequences. The number of bases compared in each case differs, owing to (a) slight differences in the starting points of the sequences (e.g., the 7N33 sequence begins at −142 instead of −196), (b) the sequences being slightly different in length due to the insertions/deletions, (c) 7N33 and 6S46 having the extra 195-bp insert, and (d) availability of published pSp2 sequence for only two-thirds of the region compared. It is clear from Figs. 3 and 5 that the sequences fall into two groups, with 6S46, 7N33, and 6S35 in one group, and pCO2 and pSp2 in the other. The 6S46, 7N33, and 6S35 sequences differ from one another by 0.34%–0.67%, pSp2 and pCO2 differ by 1.6%, but the two classes differ from one another by 2.5%–3.0%. We discuss these relationships below.

```
                5' Spacer          5'UT          Coding Sequence          3'UT          H4 - H2b Spacer
                1  2  3  4  5  6  7  8  9  10 11 12 13 14 15 A  16 17 B  18 19 20 C  D  E  21 22 23 24

6S46            .  .  T  .  .  C  C  A  C  T  T  C  G  T  A  +  A  C  +  A  .  T  +  -  -  C  .  T  T
7N33            0  0  0  0  0  C  C  A  C  T  T  C  G  T  A  =  A  C  +  A  .  T  +  -  -  C  .  T  T
6S35            0  .  T  .  .  C  C  A  C  T  T  C  G  T  C  +  A  C  +  A  A  T  -  +  -  C  .  T  T

pCO2            .  .  C  .  .  G  T  G  T  C  C  C  A  C  C  +  A  A  ----  C  G  +  -  +  T  A  C  C
pSp2            A  T  C  C  G  G  T  A  T  C  T  T  A  T  C  -  C  A  +  .  0  0  0  0  0  0  0  0  0

                25 26 27 28 29 30 F  31 32 33 34 G  35 36 37 38 H  I  39 40 41 42 43 44 45 46 47 48 49 50

6S46            T  G  C  C  TA .  +  C  CC C  C  +  T  G  A  T  -  12 T  A  C  X  T  T  A  C  T  .  C  T
7N33            T  G  C  C  NN .  +  T  TT T  T  +  T  G  A  T  -  13 T  A  C  .  T  T  A  C  T  .  C  T
6S35            A  G  C  C  .. .  -  ---------- -  T  G  A  T  -  18 T  A  A  .  T  T  A  C  T  .  C  T

pCO2            T  T  G  T  .. .  -  ---------- +  A  A  T  T  +  7  G  G  C  .  C  .  .  T  T  T  .  T
pSp2            0  0  0  0  0  C  -  ---------- +  A  A  T  G  +  5  T  G  C  T  C  .  A  T  C  N  .  .
```

**Fig. 3.** Specific sites of variation. The events indicated by symbols in Fig. 2 are summarized by listing the bases present at those sites the sequences of all five repeats. Numbers 1–50 indicate sites that involve a one- or two-base insertion/deletion or substitution. Letters A–I indicate sites of larger insertions/deletions. Pluses indicate that the sequence is present; minuses, that it is absent (Fig. 4 details these insertions/deletions). A zero is entered where base sequence is not available, and dots indicate that the base is absent from that particular clone. At site A, the equals sign refers to a third variation at the site. An entry for position 18 is not given for pCO2 because deletion B has removed the stretch of sequence containing this site. An "N" is entered if a base or bases are present but ambiguous (sites 29 and 48). An "X" is entered where the presence of any base is in question (site 42). The numbers entered in column I indicate the number of CTA repeats in the run

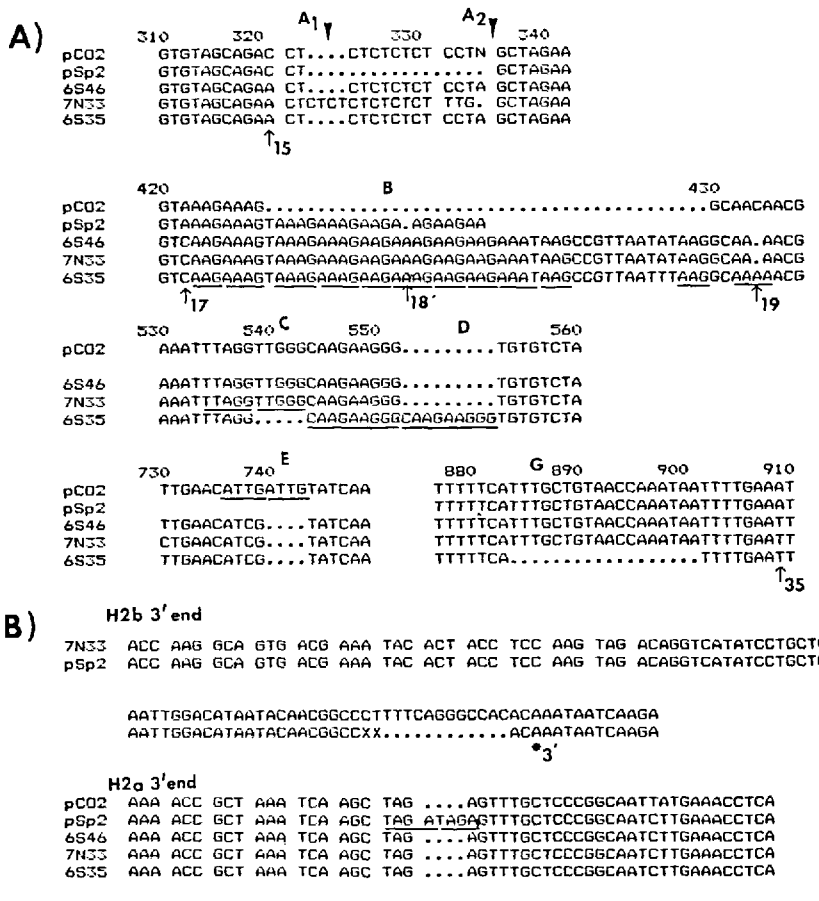


**Fig. 4A, B.** Large insertions/deletions. **A** Sequences of the larger insertions/deletions indicated in Figs. 2 and 3. Missing stretches are shown by dotted lines; repetitive stretches are underlined. Single-base events occurring in the regions are indicated by arrows. **B** Comparison of the 3′ ends of mRNA regions in *S. purpuratus* H2B and H2A genes compared in several individuals. The 3′ terminus of the H2B mRNA is indicated by an asterisk

*Locations of Mutational Events*

As indicated in Figs. 2 and 3, the region investigated consists of the H4 mRNA sequence (including the protein-coding region and the 5′ and 3′ untranslated stretches), the spacer 5′ to the H4 gene, and the spacer 3′ to the H4 gene (which we have designated the H4–H2B spacer). In two of the repeats, this spacer contains the 195-bp insert. Differences can be identified in each of these regions as indicated in Table

**Table 1.** Variation in different sequence regions

| Region | Number of bases in region[a] | Total changes[b] | Nucleotide substitutions | % total changes[c] | % nucleotide substitutions[c] |
|---|---|---|---|---|---|
| H4 5′ spacer | 127 | 6 | 2 | 4.7 | 1.6 |
| H4 5′—untranslated | 69 | 2 | 2 | 3.0 | 3.0 |
| H4 coding sequence | 306 | 6 | 6 | 2.0 | 2.0 (7.1)[d] |
| H4 3′—untranslated | 64 | 2 | 1 | 4.6 | 1.6 |
| H4–H2B spacer | 1005 | 39 | 21 | 4.0 | 2.1 |
| Insert | 195 | 4 | 3 | 2.1 | 1.5 |

[a] Number of bases in pCO2 in that particular region
[b] Changes in any clone tabulated for the respective region. Both nucleotide substitutions and insertions/deletions are counted
[c] The number of total changes or nucleotide substitutions divided by the number of bases in pCO2 in that particular region
[d] The 2.0% nucleotide substitution value for the coding region corrected for the percentage of all possible silent substitutions (Perler et al. 1980)

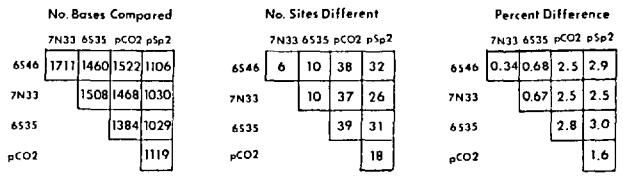


**Fig. 5.** Pairwise sequence comparisons. **Left** The number of bases compared was determined by counting the bases common to the two relevant clones in the sequence that was available. **Middle** The number of sites that differ are those sites in Fig. 3 that occur in the available sequences of the two relevant clones. **Right** Percentage difference was calculated for each pairwise comparison by dividing the number of sites that differ by the number of bases compared and multiplying by 100

1, in which each of the 61 events (at 59 sites) is given equal weight. Variation can be seen in each region: the coding sequence, nontranslated mRNA region, 5′ and 3′ spacers, and the 195-bp insert. The percentages in Table 1 are based on the lengths of the regions in pCO2 as shown in Fig. 2 (the insert is not present in pCO2).

An interesting result is that there are many base subsitutions in the coding region, all of which result in conservative (silent) codon changes. Of the total base changes possible in the coding region, 28.3% are silent changes (Perler et al. 1980). Since there is absolute conservation of the H4 protein sequence, the observed number of changes in the coding region must be normalized by the percentage of silent changes if the level of divergence in the coding region is to be compared with that of the other regions. The six nucleotide substitutions in the coding region (2.0%) correspond to a corrected value of 7.1%; it is this figure that should be used for comparison with other regions (Perler et al. 1980). The 2.0% value is almost as high as the divergences in other regions; the 7.1% level is more than twice that of any other region. As was first shown by comparison of H4 histone genes of two species (Weinberg et al. 1972; Birnstiel et al. 1974; Grunstein et al. 1975), there is apparently little selective pressure on silent site changes, even though the protein sequence is absolutely conserved.

Of the six changes noted for the 5′ spacer, five are within 40 bp of the TATA box, indicating that this region is at least as mutable as other areas surveyed. The G–C base substitution at position −91 (site 6) occurs only 7–8 bp downstream from the probable TATA box. The mutation seen in the 5′ nontranslated region at position −64 (site 7) occurs within 6 bp of the cap site. Two of the 64 base pairs in the 3′ untranslated regions were found to vary. One of the positions, site A, is a stretch of CTs that has undergone at least two kinds of change among the five repeats. One type of variation is the expansion/contraction of the CT stretch—the $(CT)_5$ of pCO2, 6S46, and 6S35 is reduced to $(CT)_1$ in pSp2 and increased to $(CT)_7$ in 7N33; the other change resulted in a replacement of CCTA by TTG in 7N33. The differences in pSp2 and 7N33 in the 3′ untranslated region would result in a mRNA length change of 16 nucleotides. We have also found insertions/deletions in the 3′ untranslated regions of other sea urchin histone genes; for example, comparison of this region from the H2B genes of 7N33 and pSp2 showed a 12-bp length difference (see Fig. 4B). It is interesting that of the 23 different histone genes examined by Hentschel and Birnstiel (1981), the pSp2 H2B sequence was the only exception to the conserved dyad symmetry in the 3′ untranslated region sequence. The 7N33 sequence in this region is virtually the same as the H2B sequence from the sea urchin *Psammechinus miliaris* and as the sequence of most other histone genes of known sequence. The difference that was noted for the H2B gene of *S. purpuratus* is therefore due to a polymorphism. We have also found length differences in the 3′ untranslated region of the *S. purpuratus* H2A gene (see Fig. 4B).

The highest percentage of variation, other than at silent codon sites, was found in the H4–H2B spacer, which contains changes in 4.0% of the po-

sitions. Only about half of this variation was due to nucleotide substitution. The values indicated in Table 1 refer to the percentage of sites that vary in any of the clones; the greatest difference in the H4–H2B spacer for any pairwise comparison is actually less than 4.0% (1.8% for base substitutions and 3% for all changes). The nucleotide substitution values in the spacer are thus lower than those for permitted silent substitutions in the coding region.

*Nature of Mutational Events*

Spacer DNA is an advantageous sequence for comparison of the relative frequencies of different kinds of mutation in a natural population. Although regions of the spacer region do have a regulatory function (Grosschedl and Birnstiel 1980a,b), only very short sequences are conserved in all sea urchin species (Hentschel and Birnstiel 1981; Maxson et al. 1983a,b). Selection at the sequence level would appear to be low, although certain short sequences (e.g., modulator regions; Grosschedl et al. 1983) or AT content might be preserved. Since there do not appear to be constraints on short insertions and deletions, the frequencies of these events can be compared with base substitutions in the spacer region. Table 2 lists the kinds of changes in the total region analyzed. The 59 positions of variation (Fig. 3) contain at least 61 different events, with sites A and 19 showing two different mutations at the same point. Our data are consistent with a Poisson distribution, which predicts that for 59 of 1711 sites undergoing at least one event, 1 or 2 sites would have two events and all other sites, one event. Therefore, we would expect the level of ambiguous second mutations (e.g., reversions or parallel substitutions) to be extremely low. Mutational site I is a special case, since there is a different number of CTA repeats in each clone; it is listed as a single event in Table 2.

Of the 61 events tabulated, there are 35 single-base substitutions and 14 single-nucleotide insertions/deletions. Two adjacent bases are substituted in 1 case and another two are inserted/deleted in another. The remaining 10 changes involve larger insertions/deletions (see Fig. 4). Of the 35 base substitutions, 21 are transitions and 14 are transversions. Transversions are rarer than would be expected if all possible base changes occur with equal probability (the expected transversions/transitions ratio would be 2:1, which from our data would have a probability of $<0.01$, $\chi^2 = 9.78$). All possible types of pairwise base changes have occurred. Since the lineage of spacer divergence can be deduced (see below), it is possible to conclude the direction of many of these changes: T to C, C to T, A to G, C to A, A to C, and T to G are found. It is also possible to determine that of the 14 single-base insertions/deletions, 5 are insertions and 2 are deletions.

**Table 2.** Types of mutational events

| | | |
|---|---|---|
| Base substitutions | | |
| Transitions | | |
| T–C | 17 | |
| G–A | 4 | |
| Total | | 21 |
| Transversions | | |
| C–A | 5 | |
| T–G | 4 | |
| C–G | 2 | |
| T–A | 3 | |
| Total | | 14 |
| Two adjacent bases | | 1 |
| Total | | 36 |
| Insertions/deletions | | |
| Single bases | 14 | |
| Two adjacent bases | 1 | |
| Larger insertions or deletions | 7 | |
| Expansion/contraction of a simple sequence | 2 | |
| Insertion of an intermediate repetitive element | 1 | |
| Total | | 25 |
| Total | | 61 |

The ten larger insertions/deletions are of several types. Event F is an insertion of a 195-bp intermediate repetitive sequence into a progenitor of 7N33 and 6S46. Event I is a family of changes resulting in fluctuations in the size of a $(CTA)_n$ stretch, with n values ranging from 5 to 18. Event A is also of this nature, resulting in a change of a $(CT)_n$ stretch, with n values from 1 to 8. Two changes are due to duplications of short sequences: CAAGAAGGG is duplicated in 6S35 (event D), and ATTG is duplicated in pCO2 (event E). Event H could be either a duplication or a deletion of the repeated sequence AGAG. Event C is a 5-bp deletion in clone 6S35 of an imperfect duplicate sequence—TTAGGTTGGG changes to TTAGG. Event B is a 33-bp deletion in pCO2 of a sequence one boundary of which occurs in the middle of an AAG or AAAG repeating sequence. Event $A_2$ may be a substitution of several bases in addition to a change in the length of the $(CT)_n$ stretch in clone 7N33. Event F (the 195-bp insertion) and event C (an 18-bp deletion in 6S35) are the only cases of larger additions/deletions not correlated with a repeated sequence.

The 195-bp insertion is found at precisely the same site in 7N33 and 6S46. It occurs between two oppositely oriented TATATAG sequences (real palindrome). There are three C–T differences and a CC–TT change between 7N33 and 6S46, clustered at one end of the insert. [The sequences of the insert will be published elsewhere (L.N. Yager, J.F. Kau-

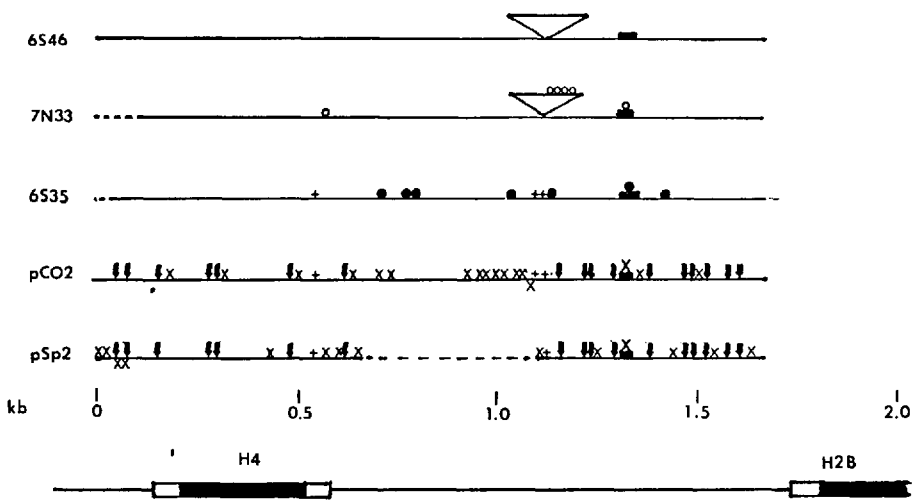


**Fig. 6.** Symbolic representation of sites shared between two or more clones. Open circles, positions shared in all repeats except 7N33, or present in the insert in 7N33 but not in 6S46; closed circles, sites shared in all repeats except 6S35; pluses, bases present in 6S35, pCO2, and pSp2, but not in 6S46 or 7N33; arrows, bases present in pCO2 and pSp2, but not in the other three repeats; *X*s, bases singular to pCO2 or pSp2, or bases singular to pCO2 in areas for which pSp2 sequence is not available. The 195-bp insert is indicated in 6S46 and 7N33. The $(CTA)_n$ stretch is indicated by a black bar

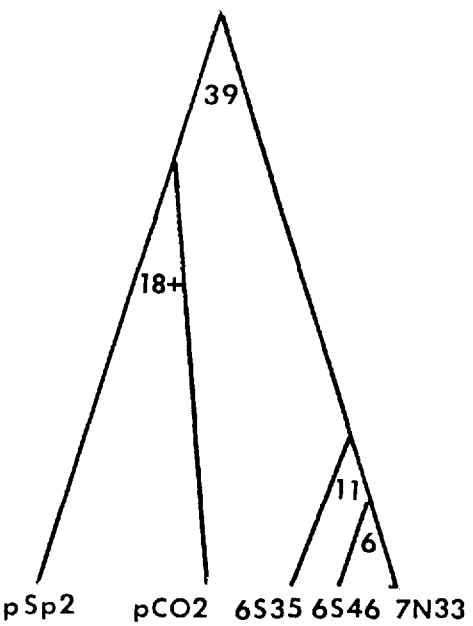


**Fig. 7.** Ancestry tree for sequence divergence. At each node the number of positions which differ between the branches is noted. The four differences within the insert are counted only for the 6S46–7N33 branchpoint. The CTA stretch is counted once at each branchpoint. The pSp2–pCO2 branchpoint is positioned arbitrarily, since the pSp2 sequence is incomplete. (Nineteen differences between pSp2 and pCO2 are found in the regions for which sequence is available for both clones)

meyer, and E.S. Weinberg, manuscript in preparation).] The inserted element occurs at 30–40 sites in every *S. purpuratus* genome, but was found within histone gene repeats of only 6 of the 22 individuals we screened.

*Spacer Lineage*

A surprising result is that the mutational events are not scattered randomly among the five sequences. The sequences clearly fall into two groups, as shown in Fig. 6. The sequence of 6S46 is used as the standard for this analysis. As shown in Figs. 3 and 5, 7N33 differs from 6S46 in only six positions: Four changes are in the insert, one is the $(CTA)_n$ stretch change, and one is the change designated $A_2$. 6S35 differs from these two very similar sequences at ten additional positions. Note that not one site is held in common between 6S35 and only pCO2 or only pSp2 (in those regions for which we have pSp2 sequence). The arrows in Fig. 6 represent changes held in common between pSp2 and pCO2 but absent from the other three repeats. The "X"s are either sites at which pCO2 and pSp2 differ from one another or are in areas in which pSp2 sequence is not available; none of these sites are changed in the other three repeats. It is likely that the "arrow class" is somewhat underrepresented, since some of the sites marked "X" in pCO2 would probably also be found in pSp2 if the missing sequence were available.

These results allow the construction of the unambiguous lineage relationship shown in Fig. 7; no other arrangement is consistent with the data. As can be seen from Fig. 3, 32 of the 59 sites were obtained from all five repeats and unambiguously require the lineage of Fig. 7. Other sites were obtained for only four of the five repeats (mostly because pSp2 sequence was not available); not only are all of these consistent with the lineage, but in most cases they require the scheme as drawn.

For this analysis $A_1$ and $A_2$ were considered two separate events (although at the same site), and the changes at position 19 were also denoted separately. The pCO2 class and the 6S46 class appear to have evolved separately for the period necessary to accumulate the various mutational differences between them. During this period, there was no recombination between the classes; thus one may gain a reasonable idea of the time frame within which

these units evolved. This subject is discussed below in further detail.

## Discussion

### *Silent Codon Changes and Distribution of Mutational Sites*

All regions of the DNA stretch we examined contain polymorphisms. Changes in the coding region are due to silent site substitutions at six positions, a level of variation as high as in the spacer regions. There is no evidence for the accumulation of silent substitutions at a rate lower than that of any other region of the repeat. A similar observation has been made recently for the alcohol dehydrogenase (Adh) gene of *Drosophila melanogaster,* in which silent sites are more polymorphic than are introns or most of the flanking sequences (Kreitman 1983). In the case of the human $\delta\beta$-globin gene region, comparison of several published sequences reveals only 2 changes within 885 bp of translated sequence (Poncz et al. 1983). The variation in 7101 bp of intron and flanking sequence is also low (0.54%), and the changes in the coding sequence are at a similar level (0.8%) when expressed as a percentage of silent changes. In these cases, at least, silent site substitutions appear to be under as little selection as is the fastest evolving intron, flanking, or spacer sequence. As pointed out previously in phylogenetic comparisons of preproinsulin and globin genes (Perler et al. 1980), silent changes are accumulated as if by a genetic clock over periods of up to 85 million years (MY) at a rate of 0.5%–1% corrected divergence per MY. For sea urchin phylogenetic comparisons, the rate of silent substitutions varies from 0.22% to 0.92% per MY (Busslinger et al. 1982). The first estimate of this type was 0.3%–0.6% per MY for the sea urchin H4 gene (Weinberg et al. 1972). Silent substitution polymorphisms may in some cases be as extensive as changes in any noncoding DNA sequence. Changes in spacer, intron, and nontranslated mRNA regions, however, would include insertion and deletion events that obviously would be highly deleterious in coding regions. The noncoding regions are more inclusive in the types of mutations observed and are therefore an excellent source of information about the relative frequencies of different types of spontaneous mutations.

Spacer sequences are probably under some selection, since particular stretches may serve regulatory (Grosschedl and Birnstiel 1980a,b; Grosschedl et al. 1983) or structural roles. Although insertions and deletions do occur, it is striking that the "early" embryonic histone gene repeats of all sea urchin species are of fairly uniform length, 5–7 kb. Even though the spacer sequences are not conserved, spacer lengths have been maintained over several hundred MY of independent evolution. The H4–H2B spacer, for example, varies only from 880 to 1350 bp in various sea urchin species (Kedes 1979). In our study, we found the distribution of mutations within the spacer to be nonrandom. If mutations are randomly distributed, the average distance between mutational sites is expected to be approximately 26 bases; however, three stretches of more than 100 bases show no variation. Two of these stretches might be expected to show no variation, since they include coding sequence (bases −37 to 72 and 114 to 228), which places restraints on replacement substitutions, but the third occurs between bases 552 and 680 in the middle of the H4–H2B spacer. Furthermore, if we consider only nucleotide substitutions, the region without polymorphisms extends from base 479 to base 704.

We do find clusters of mutations in particular regions of the histone spacer (e.g., sites 2–5, 25–30, and 43–48, according to the numbering system of Fig. 2); a certain amount of clustering would be expected by chance alone, but these areas could represent particularly mutable stretches. Site I is a mutational "hotspot": It is a run of $(CTA)_n$ differing in each sequence we studied, varying from 5 to 18 copies in the different clones. This is not a cloning artifact, since the different sizes of the $(CTA)_n$ stretches are confirmed by Southern blotting (L.N. Yager, G.C. Overton, and E.S. Weinberg, manuscript in preparation). Results indicate that the *S. purpuratus* population contains repeats with $(CTA)_n$ stretches of different length; moreover, there are only two or three sizes of $(CTA)_n$-containing fragments per individual and the hundreds of copies within a cluster are very homogeneous, with a $(CTA)_n$ length of $(n \pm 1)$ typical for each cluster.

### *Classes of Mutation*

Of the 61 mutations detected, 36 were base substitutions and 25 were insertions/deletions of various types. If we exclude the coding region, which would not accumulate frameshifts, there are 30 substitutions and 25 insertions/deletions. The insertions/deletions, therefore, are quite frequent mutational events. Of the substitutions, transitions outnumber transversions by 3:2, a ratio considerably higher than that found in the *D. melanogaster* Adh gene (Kreitman 1983), but not nearly as extreme as the 32-fold ratio of transitions to transversions found in human mitochondrial DNA (Aquadro and Greenberg 1983). The ratio obtained for the human $\delta\beta$-globin gene region, 16 transitions to 11 transversions (Poncz et al. 1983), is very similar to our value, as are the

ratios of transitions to transversions in pseudogenes (Gojobori et al. 1982). Transversions in nuclear DNA, although certainly not uncommon, thus appear to occur at a lower rate than the 2:1 excess of transversions over transitions expected if each base change were to occur with equal probability.

The most frequent type of insertion/deletion involves only one base pair. Of the 14 positions in which such an event occurred, not one was a homonucleotide run. The eight single-base-pair insertions/deletions found in the human $\delta\beta$-globin region also did not involve such runs (Poncz et al. 1983). In contrast, the few single-base insertions/deletions in the *D. melanogaster* Adh gene were all at such runs (Kreitman 1983). In the sea urchin histone genes, we did not find variation in the length of homonucleotide stretches, although the sequence of pCO2 does contain $(T)_5$, $(A)_5$, $(G)_5$, $(T)_4$, $(A)_4$, and $(G)_4$ runs. Variation in the length of a short repeat does, however, occur at sites A and I. Such expansions/contractions of short simple sequences have also been shown to be sites of polymorphism in the *D. melanogaster* Adh (Kreitman 1983), human $\delta\beta$-globin (Poncz et al. 1983), and human insulin (Ullrich et al. 1982) gene regions. A similar expansion/contraction of $(CT)_n$ occurs just 3′ to the *S. purpuratus* H2A histone gene (pCO2 versus pSp2, Kedes 1979; 7N33, L. Yager, unpublished observations). Such runs have been proposed as sites of recombination between repeat units within a histone gene cluster (Kedes 1979), but we feel that the great homogeneity of length of such runs in repeats of a single cluster makes this highly unlikely (L.N. Yager, G.C. Overton, and E.S. Weinberg, manuscript in preparation).

Of the remaining insertions/deletions, we found a single instance involving 2 bp, one insertion of an intermediate repetitive element (site F), a deletion that includes a repetitive motif (site B), and three simple duplications or deletions of short sequence (sites C, D, and E). Only one deletion, of 18 bp, is not explicable by misalignment during replication (Streisinger 1966), and no particular secondary structure (e.g., Ripley 1982) has been found that might explain this deletion.

### *Reliability of Comparisons*

The complete mRNA stretch and most of the H4–H2B spacer has been sequenced on both strands (Fig. 1). Certain regions, for example, the 3′ end of the H4–H2B spacer downstream from the Taq I site, have been sequenced on only one strand. Although these areas might be less reliable, we do find that the extent of variation, the lineage relationships, and the types of mutations found are identical to those of the bulk of the sequence. Furthermore, the three separate clones 7N33, 6S46, and 6S35 give identical sequences in these regions, as does another example of the 7N33 type sequence that has been independently cloned from a 7N library. We are therefore confident of the indicated differences between the pCO2-type and 6S35-type sequences. We did not sequence any part of pSp2, but the published sequence appears to be highly reliable. Most of the bases in pSp2 that differ from the 6S35-type sequence are also found in pCO2. In some cases a base in pSp2 differs from that in pCO2 but is found in the 6S35-type sequence. Only 11 sites are singular to pSp2; 2 of them (nos. 48 and 49) result in a change in restriction enzyme recognition sites. We have confirmed these 2 changes, and it is likely that the others are correct.

### *Evolution of the Cluster*

Although sequence polymorphisms are found in histone genes of different sea urchin individuals of a particular species, the linked repeats of a cluster are highly homogeneous (Overton and Weinberg 1978). This has now been shown at the level of direct DNA sequence by comparison of repeats from the same cluster (L.N. Yager, G.C. Overton, and E.S. Weinberg, manuscript in preparation). The evidence presented here indicates that despite the uniformity within a cluster, variation between the repeats of different individuals can be as high as 3% of the nucleotide positions. The variation, however, is not uniformly spread among all the genomes in a population. The lineage of divergence presented in Fig. 7 could only be possible if recombination between the pCO2-type and 6S35-type sequences had not occurred during the period necessary to accumulate the variation seen. The two types of sequences can be regarded as different haplotypes that have been maintained apart for several million years. We would call the lineage relationship within such nonrecombining stretches a haplotypic tree. This term would include the evolution of "frameworks" found for the human $\delta\beta$-globin genes (Orkin et al. 1982) and in tracing the sites of recombination between different haplotypes (e.g., Kreitman 1983).

There are several possible explanations for the lack of recombination in the H4 gene region; the simplest, that the individuals are from geographically isolated populations, can be ruled out, since individuals 6S and pCO2 were both collected in the Los Angeles area, whereas 7N was collected near Monterey. (The source of pSp2 is not certain.) The 6S35-type and pCO2-type sequences are therefore derived from individuals from the same area. The 195-bp insert is found in the populations in both areas, indicating gene flow between the "northern"

and "southern" populations at a time far later than that of the divergence of the two major types of sequence.

A second possibility is that the pCO2- and pSp2-type repeats are organized in minor clusters on different chromosomes from those of the major repeating class, which might preclude recombination or conversion with the 6S35-type cluster. Southern blots are not available for the individuals from which pCO2 and pSp2 were cloned, as these were derived from the DNA of several individuals many years ago. However, it is unlikely, considering the large number of copies within major clusters, that copies of minor clusters would be the first two *S. purpuratus* histone gene regions independently cloned.

The third possibility, which is the one that we favor, is that there is very little recombination or conversion between the allelic clusters during meiosis. Recombination within the tandemly repeated yeast ribosomal gene cluster has been reported to be suppressed during meiosis (Petes and Botstein 1977); there may in fact be a general suppression of meiotic recombination between gene clusters. This is in great contrast to the homogenizing events that must occur within clusters, perhaps by sister-chromatid exchange or gene-conversion events (Smith 1974, 1976; Scherer and Davis 1980). The homogenization events must be extremely rapid in comparison with recombination or gene conversion at meiosis and with the rate of mutation. No special homogenization (e.g., "molecular drive"; Dover 1982) of the repeated histone genes within the species appears to occur, since the estimated 4% divergence of single-copy DNA among sea urchin individuals (Britten et al. 1978) is similar to the divergence found in the repeated histone genes. A comparison of the sequences of specific single-copy genes from different sea urchin individuals would be informative.

We can roughly gauge the time over which the mutational events have accumulated in the 6S35- and pCO2-type repeats. Phylogenetic studies indicate that silent codon changes occur in the sea urchin early histone genes at about 0.22%–0.9% per MY (Busslinger et al. 1982). The most relevant comparison is of the "Nor 5" gene cluster of *S. drobachiensis* (there are two early gene clusters in this species) with that of *S. purpuratus*. These species diverged 4–6 MY ago and during this period accumulated 6% (corrected value) silent codon changes in the H3 gene (Busslinger et al. 1982). Changes in spacers (H2B–H3 spacers were compared) are quite a bit higher, 11.2%. The variation, based on spacer divergence, found within the *S. purpuratus* population is therefore about one-fourth that between these two species. The different *S. purpuratus* cluster types could therefore have diverged at least 1 MY ago. The silent changes, however, are as extensive within the *S. purpuratus* population as between *S. purpuratus* and *S. drobachiensis*, indicating that divergence of the cluster types may have occurred even earlier. Surprisingly, some of the nucleotide differences noted for the Nor 5 and pSp2 sequences (pSp2 was the *S. purpuratus* repeat used for the interspecies comparison) are polymorphic in *S. purpuratus*. One wonders whether many of the differences between the two species are still polymorphic in both populations. If this were true, comparisons of single units from two species would not be an accurate way to measure mutation rates and/or divergence times (Templeton et al. 1981). This factor must be considered when evaluating all previous sequence comparisons between closely related sea urchin species.

In conclusion, we have presented here the first comparison at the nucleotide sequence level of copies of reiterated genes derived from different individuals of a natural population. As expected, the variation found among these units is much greater than among units within a cluster. Pairwise comparisons of the units that we sequenced revealed that the variation was not equally distributed among different genomes. An unambiguous parsimonious tree could be constructed, indicating that the regions that we studied were evolving as haplotypes. We conclude that most of the variation among histone gene units of different clusters does not originate at meiosis by gene conversion or imprecise crossing-over or conversion events. Polymorphism is found to occur in all regions of the sequence, including the coding region. The variation within the population is fairly high, about one-fourth to one-half that found between *S. purpuratus* and *S. drobachiensis*, species that diverged some 5 MY ago. Homogenization of the histone gene repeats occurs within the cluster, but not at the population level, since much of the neutral divergence accumulated since speciation is still present. Further sequence information will be useful in learning about the evolution of the gene cluster and the history of *S. purpuratus* during its divergence from the species most closely related to it.

*Acknowledgments.* We thank Linda A. Farrelly for her excellent technical assistance. The initial cloning of 6S and 7N sequences was done by R. Donnelly. L.N. Yager was supported by NIH postdoctoral fellowship GM07799, and the project by NIH grant GM27322 to E.S.W.

## References

Aquadro CF, Greenberg BD (1983) Human mitochondrial DNA variation and evolution: analysis of nucleotide sequences from seven individuals. Genetics 103:287–312

Benton WD, Davis RW (1977) Screening λgt recombinant clones by hybridization to single plaques in situ. Science 196:180–182

Birnstiel ML, Weinberg ES, Pardue ML (1974) Evolution of 9S mRNA sequences. In: Hamkalo BA, Papaconstantinou J (eds) Molecular cytogenetics. Plenum Press, New York London, pp 75–93

Blattner FR, Williams BG, Blechl AE, Denniston-Thompson K, Faber HE, Furlong L, Grunwald DJ, Kiefer DO, Moore DD, Schumm JW, Sheldon EL, Smithies O (1977) Charon phages: safer derivatives of bacteriophage lambda for DNA cloning. Science 196:161–169

Britten RJ, Cetta A, Davidson EH (1978) The single copy DNA sequence polymorphism of the sea urchin *Strongylocentrotus purpuratus*. Cell 15:1175–1186

Brutlag DL, Clayton J, Friedland P, Kedes LH (1982) SEQ: a nucleotide sequence analysis and recombination system. Nucleic Acids Res 10:279–294

Busslinger MD, Rusconi S, Birnstiel ML (1982) An unusual evolutionary behaviour of a sea urchin histone gene cluster. EMBO J 1:27–33

Childs G, Nocente-McGrath C, Lieber T, Holt C, Knowles J (1982) Sea urchin (*Lytechinus pictus*) late-stage histone H3 and H4 genes: characterization and mapping of a clustered but nontandemly linked multigene family. Cell 31:383–393

Dover G (1982) Molecular drive: a cohesive mode of species evolution. Nature 299:111–117

Gojobori T, Li W-H, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360–369

Grosschedl R, Birnstiel ML (1980a) Identification of regulatory sequences in the prelude sequences of an H2A histone gene by study of specific deletion mutants *in vivo*. Proc Natl Acad Sci USA 77:1432–1436

Grosschedl R, Birnstiel ML (1980b) Spacer DNA sequences upstream of the TATAAAATA sequence are essential for promotion of H2A histone gene transcription *in vivo*. Proc Natl Acad Sci USA 77:7102–7106

Grosschedl R, Machler M, Rohrer U, Birnstiel ML (1983) A functional component of the sea urchin H2A gene modulator contains an extended sequence homology to vival enhancer. Nucleic Acids Res 11:8123–8136

Grunstein M, Schedl P, Kedes L (1975) Sequence analysis and evolution of sea urchin (*Lytechinus pictus* and *Strongylocentrotus purpuratus*) histone H4 messenger RNAs. J Mol Biol 104:323–349

Grunstein M, Diamond KE, Koppel E, Grunstein J (1981) Comparison of the early histone H4 gene sequence and late histone H4 mRNA sequences. Biochemistry 20:1216–1223

Hentschel C, Birnstiel M (1981) The organization and expression of histone gene families. Cell 25:301–313

Hohn B, Murray K (1977) Packaging recombinant DNA molecules into bacteriophage particles *in vitro*. Proc Natl Acad Sci USA 74:3259–3263

Joseph DR, Stafford DW (1976) Purification of sea urchin ribosomal RNA genes with a single-strand specific nuclease. Biochim Biophys Acta 418:167–174

Kedes LH (1979) Histone genes and histone messengers. Annu Rev Biochem 48:837–870

Krietman M (1983) Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature 304: 412–417

Maniatis T, Fritsch EF, Sambrook J (1982) In: Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp 117–119

Mauron A, Levy S, Childs G, Kedes LH (1981) Monocistronic transcription is the physiological mechanism of sea urchin embryonic histone gene expression. Mol Cell Biol 1:661–671

Maxam AH, Gilbert W (1980) Sequencing end labeled DNA with base-specific chemical cleavages. Methods Enzymol 65: 499–560

Maxson R, Mohun T, Cohn R, Kedes L (1983a) Expression and organization of histone genes. Annu Rev Genet 17:237–277

Maxson R, Mohun T, Gormezano G, Childs G, Kedes L (1983b) Distinct organizations and patterns of expression of early and late histone gene sets in the sea urchin. Nature 301:120–125

Orkin SH, Kazazian HH Jr, Antonarakis SE, Goff SC, Boehm CD, Sexton JP, Waber PG, Glardina PJV (1982) Linkage of $\beta$-thalassaemia mutations and $\beta$-globin gene polymorphisms with DNA polymorphisms in human $\beta$-globin gene cluster. Nature 296:627–631

Overton C, Weinberg E (1978) Length and sequence heterogeneity of the histone gene repeat unit of the sea urchin, *S. purpuratus*. Cell 14:247–257

Perler F, Efstratiadis A, Lomedico P, Gilbert W, Kolodner R, Dodgson J (1980) The evolution of genes: the chicken preproinsulin gene. Cell 20:555–566

Petes TD, Botstein D (1977) Simple Mendelian inheritance of the reiterated ribosomal DNA of yeast. Proc Natl Acad Sci USA 74:5091–5095

Poncz M, Schwartz E, Ballantine M, Surrey S (1983) Nucleotide sequence analysis of the $\beta$-globin gene region in humans. J Biol Chem 258:11599–11609

Ripley LS (1982) Model for the participation of quasipalindromic DNA sequences in frameshift mutation. Proc Natl Acad Sci USA 79:4128–4132

Scherer S, Davis RW (1980) Recombination of dispersed repeated DNA sequences in yeast. Science 209:1380–1384

Smith GP (1974) Unequal crossover and the evolution of multigene families. Cold Spring Harbor Symp Quant Biol 38: 507–513

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191:528–535

Stafford DW, Bieber D (1975) Concentration of DNA solutions by extraction with 2-butanol. Biochim Biophys Acta 378: 18–21

Streisinger G (1966) Frameshift mutations and the genetic code. Cold Spring Harbor Symp Quant Biol 31:77–84

Sures I, Lowry J, Kedes LH (1978) The DNA sequence of sea urchin (*S. purpuratus*) H2A, H2B and H3 histone coding and spacer regions. Cell 15:1033–1044

Sures I, Levy S, Kedes LH (1980) Leader sequences of *Strongylocentrotus purpuratus* histone mRNAs start at a unique heptanucleotide common to all five histone genes. Proc Natl Acad Sci USA 77:1265–1269

Templeton AR, DeSalle R, Walbot V (1981) Speciation and inferences on rates of molecular evolution from genetic distances. Heredity (Edinburgh) 47:439–442

Thomas M, Cameron JR, Davis RW (1974) Viable molecular hybrids of bacteriophage lambda and eukaryotic DNA. Proc Natl Acad Sci USA 71:4579–4583

Ullrich A, Dull TJ, Gray A, Phillips JA, Peter S (1982) Variation in the sequence and modification state of the human insulin gene flanking regions. Nucleic Acids Res 10:2225–2240

Weinberg ES, Birnstiel ML, Purdom IF, Williamson R (1972) Genes coding for polysomal 9S RNA of sea urchins: conservation and divergence. Nature 240:225–228